Presented at ESCA workshop of Robust Speech Recog., Pont-a-Mousson (1997), pp. 87-90.

IMPROVING ASR PERFORMANCE FOR REVERBERANT SPEECH

Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg
email: {bedk,morgan,steveng}@icsi.berkeley.edu
International Computer Science Institute
1947 Center St., Suite 600, Berkeley, CA 94704
USA

ABSTRACT

The performance of current automatic speech recognition (ASR) systems is very sensitive to the presence of room reverberation in the incoming speech signal. We investigate a family of front-end speech representations that focus on slow changes in the the gross spectral structure of speech for their ability to improve the robustness of ASR systems to reverberation. A number of the front ends provide a statistically significant improvement in performance over established front ends such as PLP; however, the performance of ASR systems on highly reverberant speech is still disappointing when compared with the performance of human listeners.

1. INTRODUCTION

The problem of reliably recognizing reverberant speech is an important one for automatic speech recognition systems. In virtually every application in which the use of head-mounted, close-talking microphones is excluded, room reverberation can significantly alter incoming speech. As earlier studies have shown (see below), current recognizers cannot reliably recognize even slightly reverberant speech. Theoretically, reverberation constitutes a form of distortion that is distinct from both short-time spectral coloration and additive noise. Although it is a form of convolutional distortion, reverberation, unlike short-time spectral coloration, is not multiplicative in the short-time spectral domain because the duration of room impulse responses is typically longer than the temporal window used for spectral analysis of speech. Reverberation instead appears as a form of temporal smearing in the time-frequency plane. Unlike additive noise, the distortion imposed by room reverberation is completely correlated with the speech signal.

These differences make it likely that conventional approaches to robust speech recognition, which have typically been developed on data corrupted by spectral coloration, additive noise, or both, will not be effective in the presence of reverberation. For phone recognition on TIMIT sentences, it is reported in [1] that the phone error rate of a recognizer using a mel-cepstral front end increases from 27.1% on a clean test set to 81.3% on a test set processed through a room reverberation simulator with a reverberation time of roughly 250–300 ms. A recognizer that uses an auditory-based front end that is robust to additive noise, the ensemble interval histogram (EIH), has a phone error rate of 36% on the clean test set and an error rate of 82.7% on the reverberant test set. Recently, we have compared the performance[1] of ASR systems using the PLP [4], log-RASTA-PLP [5], and J-RASTA-PLP [5] front ends with the performance of human listeners on a highly reverberant test set [2]. The results of this study are summarized in Table 1. Generally, the word error rate for the ASR systems increases by a factor of four when going from the clean test set to the reverberant test set, and on the reverberant test set the ASR systems' word error rate is roughly a factor of twelve higher than the word error rate of human listeners. Clearly, while humans are adept at recognizing even highly reverberant speech, ASR systems are not.

	clean error	reverb. error
PLP	15.8%	70.1%
log-RASTA-PLP	14.5%	72.7%
J-RASTA-PLP	15.1%	77.3%
humans	—	6.1%

Table 1. Word error rates for three different ASR front ends and for human listeners. The task, which is described in more detail in Section 3, is word recognition on connected strings of numbers. The reverberant test set is generated by convolving the clean test set with an impulse response that has a roughly 2 s reverberation time in the mid-frequencies. Differences of 2% word error rate or greater are statistically significant ($p < 0.05$). Note that the human listeners recognize the reverberant speech more reliably than the ASR systems recognize the clean speech.

How might ASR performance on reverberant speech be improved? We believe that one important step is to use a representation of speech that focuses on slow modulations distributed across critical-band-like channels. In the rest of this paper we justify the use of such a representation, describe the design of and results from a series of experiments that explore the use of modulation-based representations for ASR in reverberation, and discuss the implications of these results.

2. THE IMPORTANCE OF SLOW MODULATIONS FOR SPEECH INTELLIGIBILITY

Ideally, a representation of speech used for recognition should focus on those aspects of the speech signal that encode phonetic information and should suppress those aspects that do not. A considerable body of evidence points to the importance of slow changes in the gross speech spectrum, which appear as modulations of energy at rates of 2–16 Hz in roughly critical-band-wide channels, for conveying the phonetic information in the speech signal. As early as the late 1930's, the developers of the vocoder found that they could synthesize high-

[1] It should be noted that the recognition scores reported here are slightly different from those reported in [2] and [3]. Since the publication of these two papers, we discovered and corrected a small systematic error in the scoring of our recognition results. The numbers reported here are the correct recognition scores.

experiment	envelope norm.	complex filter	global peak	output thresh.	clean error	reverb. error
A	X	X	X	X	30.1%	65.2%
B	X		X	X	30.6%	67.8%
C	X	X	X		17.5%	66.1%
D	X		X		13.6%	69.9%
E		X			18.3%	68.8%
F					16.1%	73.5%

Table 2. Summary of the recognition results for different variants of the modulation spectrogram front end. An "X" in a column indicates that the corresponding processing step was used in the front end. Thus, the first line of the table is the baseline modulation spectrogram result. Differences of 2% word error rate or greater are statistically significant ($p < 0.05$).

quality speech using a dynamic estimate of spectral shape that was low-pass filtered at 25 Hz [6]. More recently, the speech transmission index [7], a measure that summarizes the low-frequency modulation transfer function of a channel, has proven to be a powerful predictor of the intelligibility of speech transmitted via a wide variety of channels, including reverberant and noisy rooms [8]. Finally, in a recent set of experiments on the intelligibility of temporally-smeared speech (which is very similar to reverberant speech), Drullman and colleagues have shown that modulations at rates above 16 Hz are not required for speech intelligibility [9].

We already have two front-end speech representations that focus on slow modulations in speech, log-RASTA-PLP and J-RASTA-PLP, that have been successfully used to enhance the robustness of ASR systems to spectral coloration and (with J-RASTA-PLP) additive noise. Thus, the results in Table 1 are surprising. Both log-RASTA-PLP and J-RASTA-PLP are outperformed by PLP, a front end that does no processing of modulations, on the reverberant test set. If the principle of focusing on slow modulations is correct, then there must be some detail of the processing performed by the RASTA front ends, the domain in which modulation filtering is performed, for example, that makes them unsuitable for recognizing reverberant speech.

We have developed a new representational format for speech, the modulation spectrogram [3], that displays the distribution of slow modulations across time and frequency. To produce a modulation spectrogram, a speech signal is processed through the following steps:

1. The speech signal, which in our application is sampled at 8 kHz, is analyzed into critical-band-like subbands. A fixed quarter-octave bandwidth is used in a fifteen-channel FIR filter bank that covers 297–4000 Hz. The filter transfer functions are trapezoidal, with minimal overlap between adjacent channels.
2. The amplitude envelope in each subband is computed by full-wave rectification and low-pass filtering with a cutoff frequency of 20 Hz. The envelope signal is downsampled to 80 Hz.
3. In each subband, the average envelope level is computed over an entire utterance, and the envelope signal is divided by this average level. This normalization functions as a form of automatic gain control, suppressing any spectral coloration of the speech signal and ensuring that the signals in all of the subbands have roughly equal levels.
4. In each subband, a complex FIR filter is used to estimate the spectral power of modulations between 0–8 Hz. The log (base ten) of the squared magnitude of the filter output is taken.
5. The maximum modulation level for all bands over an entire utterance is found, and then is subtracted from all of the subband modulation signals. This step also functions as a form of automatic gain control, ensuring that the peak output level for any utterance is 0 dB.
6. Thresholding is applied to all of the modulation signals, so levels that are below -30 dB (referenced to the global maximum level) are set to -30 dB. This restricts the dynamic range of the representation to 30 dB, and suppresses lower-energy portions of the signal.

It should be noted that the details of the processing used in this paper are slightly different than the processing described in [3]; however, these differences do not greatly affect the final representation.

The modulation spectrogram was initially developed to produce visual displays of speech that are stable in low signal-to-noise ratio and in highly reverberant conditions, and not as a representation of speech for ASR. We were therefore interested in addressing a number of questions. First, could the same processing that was used to produce the visual displays be used for reverberation-robust ASR? Second, could the processing be better tuned for use in ASR systems? Third, what steps in the processing are most important for improving the robustness of ASR systems to reverberation? Fourth, can we explain the poor performance of the RASTA front ends on reverberant speech?

3. MATERIALS AND METHODS

We attempted to answer these questions by running experiments in which an ASR system was trained on a clean set of training utterances, using different variants on the basic modulation spectrogram described above as the front-end processing. The performance of the ASR system, in terms of word error rate, was then measured on clean and reverberant versions of a test set. The results for the clean test set indicate how well the front end preserves phonetic information in the speech signal, while the results for the reverberant test indicate how robust the representation of phonetic information is to reverberation.

All experiments were performed using material from Numbers93, a subset of the Numbers corpus collected by the Center for Speech and Language Understanding at the Oregon Graduate Institute. Numbers is a collection of spontaneous utterances collected over the telephone from a diverse population of speakers. The utterances are digitized at an 8 kHz sampling rate with a 16-bit A/D converter. The vocabulary of the subset we used is restricted to numbers (including confusable sets like "seven," "seventy," and "seventeen") and a few other words (e.g. "oh," "double," and "and"). The training set contains 875 utterances, while the test set contains 657 utterances.

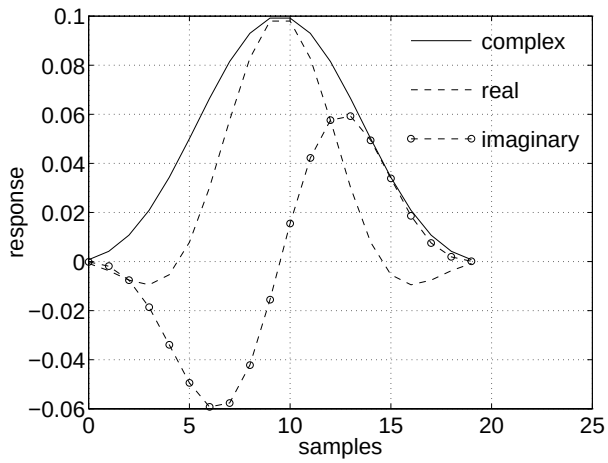

Figure 1. A comparison of the response of the complex filter with the responses of its real and imaginary components. The response of the complex filter is computed as the square root of the sum of the squares of the responses of the real and imaginary components. Note that the response of the complex filter is broader than that of either the real or the imaginary component.

The reverberant test set was generated by digitally convolving the utterances in the clean test set with a hand-designed impulse response that was intended to match the characteristics of a hallway about 6.1 m long, 2.4 m high, and 1.7 m wide, with concrete walls, floor, and ceiling. The reverberation time of the hallway varies from a low of 1.4 s in the 2–4 kHz band to a high of 3.1 s in the 0–250 Hz band. The ratio of direct to reverberant energy in the impulse response is -16 dB. Details of the design of the impulse response are presented in [2].

The ASR system used in the experiments is a hybrid hidden Markov model/multilayer perceptron (HMM/MLP) recognizer [10] that uses an MLP to estimate phone probabilities from acoustic features and uses Viterbi search for speech decoding. The modulation spectrogram front ends tested generate output frames at a rate of 80 Hz, and usually generate fifteen output values per frame. The MLP phone probability estimator takes the current output of the front end, the previous seven frames, and the next seven frames as input, and thus typically has 225 input units. The MLP has 56 output units, and typically 320 hidden units. The total number of weights in the MLP was fixed at 89920, so for tests in which the front end produced more that 15 outputs per frame, the number of units in the hidden layer is reduced to keep the total number of weights constant. The speech decoder uses a single-pronunciation lexicon and a class bigram grammar language model. For recognizer training, an iterative procedure that trains a recognizer to an initial labeling of the training set, then relabels the data via forced alignment and trains a new recognizer on the new labels was used to ensure a good match between the features and the word models used for recognition.

4. EXPERIMENTS

In the first set of experiments, we tested the modulation spectrogram as described above and variants in which different processing steps were omitted. Steps that were optionally left out are the normalization of the subband envelope signals by the average envelope levels, the complex filtering that measures modulation level in the 0–8 Hz range, the normalization of the global peak to a level of 0 dB, and the thresholding of levels below -30 dB. Note that if all of these steps are omitted, the output of the front end is the log of the squared subband envelope signals. Table 2 summarizes the results of these experiments.

filter	compression	clean error	reverb. error
complex	log	17.8%	63.8%
complex	cube root	17.8%	67.2%
real	cube root	16.5%	68.3%
imaginary	cube root	17.3%	64.3%
real and imag.	cube root	14.7%	63.5%

Table 3. Summary of the recognition results for different modulation filters. Differences of 2% word error rate or greater are statistically significant ($p < 0.05$).

A number of interesting patterns are evident in these results:

- The basic modulation spectrogram features (experiment A), which produce visual displays that are stable in low signal-to-noise ratio and in highly reverberant conditions, are not adequate for use as an ASR front end, at least for recognizing clean speech. The performance on the reverberant test set is good compared to the other front ends, although it is a factor of ten worse than that of human listeners. It should be noted that a recognizer that uses phone probabilities from an MLP trained on these features in combination with probabilities from an MLP trained on PLP features has a 13.6% word error rate on the clean test set and a 64.1% word error rate on the reverberant test set [2].
- If the results of experiments A and C are compared, it is clear that the thresholding that is needed to produce stable visual displays causes problems for ASR systems. It is likely that phonetic information carried by low-energy segments is destroyed by the thresholding. The recognizer trained in experiment C has performance on the clean test set that is not significantly different from the performance of a recognizer trained on PLP features (see Table 1) and performance on the reverberant test set that is not significantly different from the performance of the recognizer in experiment A. It is possible that the use of a lower threshold value, for example -60 dB instead of -30 dB, would also give good results for clean speech. It is also possible that thresholding, which does not appear to be effective on the highly reverberant test set we used, would be more useful on moderate levels of reverberation. We have not yet investigated either of these questions.
- If the results of pairs of experiments in which the only difference in the front end is the presence or absence of the complex filter (experiments A and B, C and D, and E and F) are compared, it is apparent that the filtering is an important factor for improving recognition in reverberation. In all cases, omission of the filter results in a statistically significant decrease in recognition performance on the reverberant test set.
- The best performance on the clean test set is obtained in experiment D, where only the envelope normalization and referencing of the output to the global peak value are performed. It is likely that these steps suppress any channel effects in the data, which was collected over telephone lines, and thus improve recognition performance for the clean test set.

In a second set of experiments we examined the filter used to detect slow modulations. While use of a complex

filter to estimate the spectral power of slow modulations is convenient for a number of reasons (e.g. the magnitude of the output of the spectral filter is never negative, so the output is convertible to dB), there are also some drawbacks. The complex filter requires twice as much computation as a real filter. Also, the temporal response of the complex filter is broader than the response of either the real or the imaginary component of the filter (see Figure 1). We therefore investigated the possibility of replacing the complex filter with either its real or imaginary component, or with both. Because the outputs of the real and imaginary components may be negative, the log compression on the modulation filter output was replaced with a cube root. These features were computed without the envelope signal normalization and without thresholding of the output. Table 3 summarizes the results of these experiments, and compares them to the performance of a modulation spectrographic front end that uses the complex filter with logarithmic compression of the output, no envelope normalization, and no thresholding.

Changing the compression from logarithmic to cube root causes a significant degradation in performance on the reverberant test set when the complex filter is used. This change in performance may be due to the lesser amount of compression imposed by the cube root, compared with the logarithm, at high signal levels. With the cube root compression, there is no significant difference in performance on the clean test set for either the complex filter, its real component, or its imaginary component. However, using the imaginary component provides a significant performance improvement over the complex filter or its real component on the reverberant speech. This improvement is probably due to the imaginary component's enhancement of changes in the envelope signal. Finally, using both the real and imaginary components instead of combining them into a single magnitude gives the best overall performance. This case is analogous to using both features and delta features, which are essentially orthogonal representations of the speech, for recognition. The real component of the filter is a smoothing filter, while the imaginary component is a differentiator.

5. DISCUSSION

Several of the representations derived from the basic modulation spectrogram perform as well as PLP on the clean test set and significantly better than PLP on the reverberant test set. The key processing step that produces this robustness to reverberation is the modulation filtering. By focusing the recognizer's modeling power on the slow modulations in speech, it is possible to improve performance in reverberant conditions. Furthermore, it appears that the domain in which this filtering is done may indeed be important. While we are able to improve recognizer performance with a front end that filters modulations in the linear domain, filtering modulations in the logarithmic or linear-logarithmic domains, as in log-RASTA-PLP and J-RASTA-PLP respectively, degrades recognizer performance on reverberant speech.

The experiments which use the imaginary component of the original complex modulation filter suggest that performance in reverberation may be enhanced by using a bandpass modulation filter instead of a lowpass filter. The imaginary component of the complex filter is actually a bandpass filter with a passband covering 1.5–8 Hz. This result is consistent with the results in [11], where highpass filtering of envelope signals was used to improve the reverberation robustness of an isolated-word dynamic-time-warp recognizer.

Although we are able to improve the robustness of our ASR system to reverberation by using a modulation-based front end instead of PLP, the performance of the recognizer is still disappointingly poor in comparison to the performance of human listeners. Moreover, in [2] we found that even when the training and testing conditions were matched (the recognizer was trained on a reverberant version of the training set then tested on the reverberant test set), performance was still poor. A recognizer with a PLP front end had a word error rate of 48.5% in this experiment, while a recognizer with a modulation spectrogram front end had a word error rate of 43.5%. This result indicates that the performance of ASR systems in highly reverberant conditions may be limited by more than simply the front-end signal processing [12].

REFERENCES

[1] Sumeet Sandhu and Oded Ghitza. A comparative study of mel cepstra and EIH for phone classification under adverse conditions. In *ICASSP-95. The 1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412. IEEE, 1995.

[2] Brian E. D. Kingsbury and Nelson Morgan. Recognizing reverberant speech with RASTA-PLP. In *ICASSP-97. 1997 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1997.

[3] Steven Greenberg and Brian E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP-97. 1997 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1997.

[4] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[5] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[6] Homer Dudley. Remaking speech. *Journal of the Acoustical Society of America*, 11(2):169–177, October 1939.

[7] Tammo Houtgast and Herman J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility. *Journal of the Acoustical Society of America*, 77(3):1069–1077, March 1985.

[8] H. J. M. Steeneken and T. Houtgast. Evaluation of a physical method for estimating speech intelligibility in auditoria. In *ICASSP 82. 1982 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1452–1454. IEEE, 1982.

[9] Rob Drullman, Joost M. Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95(2):1053–1064, February 1994.

[10] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[11] H. G. Hirsch. Automatic speech recognition in rooms. In J. L. Lacoume, A. Chehikian, N. Martin, and J. Malbos, editors, *Signal Processing IV: Theories and Applications. Proceedings of EUSIPCO-88. Fourth European Signal Processing Conference.*, volume 3, pages 1177–1180, Amsterdam, 1988. Elsevier Science Publishers B.V.

[12] Steven Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Channels*. ESCA, 1997.